# Exact protein distributions for stochastic models of gene expression using partitioning of Poisson processes

Pendar, H. , Platini, T. and Kulkarni, R.V.

# Exact protein distributions for stochastic models of gene expression using partitioning of Poisson processes

Hodjat Pendar*
*Department of Engineering Science and Mechanics, Virginia Tech, Blacksburg, Virginia 24061, USA*

Thierry Platini†
*Applied Mathematics Research Center, Coventry University, Coventry, CV1 5FB, England*

Rahul V. Kulkarni‡
*Department of Physics, University of Massachusetts, Boston, Massachusetts USA*

(Received 16 February 2013; revised manuscript received 4 April 2013; published 26 April 2013)

Stochasticity in gene expression gives rise to fluctuations in protein levels across a population of genetically identical cells. Such fluctuations can lead to phenotypic variation in clonal populations; hence, there is considerable interest in quantifying noise in gene expression using stochastic models. However, obtaining exact analytical results for protein distributions has been an intractable task for all but the simplest models. Here, we invoke the partitioning property of Poisson processes to develop a mapping that significantly simplifies the analysis of stochastic models of gene expression. The mapping leads to exact protein distributions using results for mRNA distributions in models with promoter-based regulation. Using this approach, we derive exact analytical results for steady-state and time-dependent distributions for the basic two-stage model of gene expression. Furthermore, we show how the mapping leads to exact protein distributions for extensions of the basic model that include the effects of posttranscriptional and posttranslational regulation. The approach developed in this work is widely applicable and can contribute to a quantitative understanding of stochasticity in gene expression and its regulation.

DOI: 10.1103/PhysRevE.87.042720 PACS number(s): 87.10.Mn, 02.50.−r, 82.39.Rt, 87.17.Aa

## I. INTRODUCTION

One of the fundamental problems in biology is the elucidation of molecular mechanisms that give rise to phenotypic variations among individuals in a population. Recent research has shown that phenotypic variations can arise without any underlying differences in the genotype or environmental factors [1,2]. Such "nongenetic individuality" is driven by fluctuations (noise) in cellular levels of gene expression products, as observed in diverse processes ranging from bacterial persistence [3] to HIV-1 viral infections [4]. Quantifying and modeling noise in gene expression is thus an important step toward a fundamental understanding of phenotypic variation among genetically identical cells.

Noise in gene expression is generally analyzed using coarse-grained stochastic models [5,6]. For such models, cellular variations can be characterized using the mean and variance of mRNA and protein distributions [6–9]. However, in several cases, it is of interest to characterize the entire distribution, rather than just the mean and variance. For example, it has been demonstrated that protein distributions can exhibit features such as bimodality [10] that are not adequately represented using the first two moments alone. Since protein levels in single cells can be measured experimentally [11,12], developing analytical approaches for protein distributions is an important contribution toward building quantitative models of gene expression.

Given the need for analytical results for the entire distribution, several approaches have been developed in recent work. Analytical results for mRNA distributions have been derived [13–19]; however, the corresponding results for proteins have been significantly more challenging to obtain. When the mean mRNA lifetimes ($\tau_m$) are much shorter than protein lifetimes ($\tau_p$), analytical expressions have been derived for protein steady-state distributions [20,21]. More generally, exact results have recently been derived [22] for the simplest model of gene expression, also known as the two-stage model. While useful results have thus been obtained, further generalizations are needed to include a broader class of models that include the effects of cellular regulation.

In this paper, we develop an analytical framework that leads to exact protein distributions for a wide range of stochastic models of gene expression. In the following section, we provide brief definitions of some basic concepts used in the analysis.

## II. MASTER EQUATION AND GENERATING FUNCTIONS

Defining the probability distribution $\Phi(X,t)$ to find the system under consideration in a given state $X$ at a time $t$, the corresponding master equation is given by

$$\partial_t \Phi(X,t) = \sum_Y \left[\Phi(Y,t)w_X^Y - \Phi(X,t)w_Y^X\right], \qquad (1)$$

where $w_Y^X$ is the rate of transition from $X$ to $Y$.

It is often the case that the state of the system ($X$) is fully characterized by a set of integers ($\{n_j\}$) such as the number of mRNA, proteins, etc. It follows that the probability distribution

*Electronic address: hpendar@vt.edu
†Electronic address: thierry.platini@coventry.ac.uk
‡Electronic address: rahul.kulkarni@umb.edu

 

becomes $\Phi(\{n_j\},t)$. The corresponding generating function $G$ (a function of a set of continuous variables $\{x_j\}$) is defined by

$$G(\{x_j\},t) = \sum_{\{n_j\}} x_1^{n_1} x_2^{n_2} \ldots x_q^{n_q} \Phi(\{n_j\},t). \quad (2)$$

All the moments of the probability distribution $\Phi(\{n_j\},t)$ can be obtained from $G$ by succesive differentiation. Finally, the entire probability distribution can also be obtained from the expression for $G$, either analytically or by using numerical approaches. In the following, we develop an analytical framework for obtaining the generating function $G$ for protein distributions in stochastic models of gene expression.

## III. MAPPING TO REDUCED MODELS

We will consider models of gene expression for which the creation of mRNAs is a Poisson process occurring with rate $k_m$. Invoking a well-known theorem on the partitioning of Poisson processes [23], we develop a mapping that significantly simplifies analysis of such models.

We begin by partitioning the mRNA arrivals into $N$ "types" [Fig. 1(a)]. Given a mRNA arrival at any time $t$, the probability that it is assigned to type $i$ $(i = 1 \ldots N)$ is $q_i = 1/N$. Thus, each mRNA is equally likely to be assigned to one of the $N$ types upon arrival. Denoting by $\mathcal{N}_i(t)$ the number of arrivals of the $i$th type of mRNA by time $t$, it follows from the theorem of partitioning of Poisson processes [23], that the arrival of each type of mRNA is an independent Poisson process occurring with rate $k_m/N$ [Fig. 1(a)]. In other words, the $\mathcal{N}_i(t)$ $(i = 1 \ldots N)$ are independent Poisson random variables with mean $\langle \mathcal{N}_i(t) \rangle = k_m t/N$.

The next step consists of taking the limit $N \to \infty$ and leads to the definition of the reduced model. For any given time $t$, in the limit $N \to \infty$, the probability of arrival of more than one mRNA of any given type can be neglected (see Appendix A).

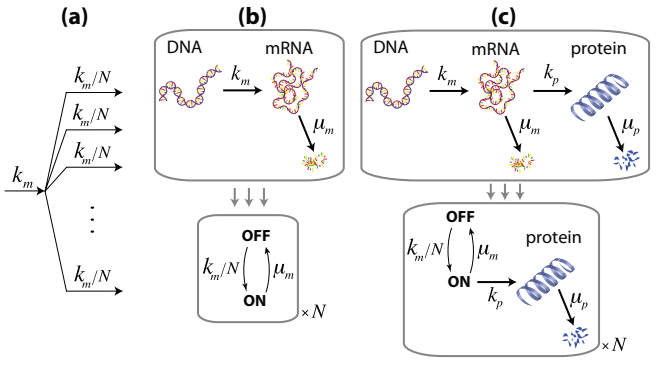


FIG. 1. (Color online) (a) A Poisson arrival process with arrival rate $k_m$ can be partitioned to $N$ independent and identical Poisson arrival processes, each occurring with rate $k_m/N$. (b) Partitioning of the Poisson arrival process leads to a mapping from a simple model of creation and decay of mRNAs to $N$ independent, identical two-state systems (in the limit $N \to \infty$). The probability of having $m$ mRNAs in the original model is equivalent to the probability of having $m$ two-state systems in the "ON" state in the reduced model. (c) The same mapping applied to the two-stage model of gene expression for proteins. Note that the reduced model is identical to a model for creation and decay of mRNAs with promoter-based regulation.

It follows that the random variable describing the number of mRNAs of a given type is constrained to the value 0 or 1. Effectively, after partitioning of the Poisson arrival process, the mRNA dynamics can be replaced by the dynamics of a two-state system. Thus, at any time $t$, we have a mapping from the original system to $N$ identical subsystems. In the limit $N \to \infty$, each of these subsystems corresponds to what will be referred to as a "reduced" model. Further details on the connection between original and reduced models is provided in Appendix A. In the following, we will refer to this approach as the PPA (partitioning of Poisson arrivals) mapping.

As an illustration, let us consider the number of mRNAs for the simple model shown in Fig. 1(b). It is readily derived (e.g., using the Master equation) that the corresponding steady-state distribution is a Poisson distribution with mean $k_m/\mu_m$. This result can also be obtained using the PPA mapping, as illustrated in Fig. 1(b). The dynamics of the reduced model (a two-state model) is defined by the transitions between 0 mRNA ("OFF") and 1 mRNA ("ON") states driven by the rates $k_m/N$ and $\mu_m$. Therefore, the steady-state generating function for mRNAs in the reduced model is given by $g(z) = (1 - \frac{k_m/N}{\mu_m + k_m/N}) + \frac{k_m/N}{\mu_m + k_m/N} z$. Correspondingly, the generating function for the distribution of mRNAs in the original model is given by $G(z) = \lim_{N\to\infty}[g(z)]^N$. This expression reduces to the generating function of the Poisson distribution with mean $k_m/\mu_m$, thereby recovering the well-known result. An explicit derivation illustrating this approaching using the master equation is provided in Appendix B.

The preceding argument can be generalized to analyze the distribution of proteins in stochastic models of gene expression. In order to apply the PPA mapping, we will consider models for which the protein production from each mRNA proceeds independently. Let $P(t)$ be the random variable corresponding to the number of proteins in the system at time $t$. Partitioning the mRNAs into $N$ "types," we denote by $p_i$ the random variable corresponding to the number of proteins created by the $i$th type of mRNA. Note that, in the limit $N \to \infty$, $p_i$ is the random variable corresponding to the distribution of proteins in the reduced model. Since each mRNA contributes independently, the $p_i(t)$ are independent, identically distributed random variables such that $P = \sum_{i=1}^{N} p_i$. Correspondingly, the generating functions for proteins in the original $[G(z,t)]$ and reduced $[g(z,t)]$ models are related by

$$G(z,t) = \lim_{N\to\infty} [g(z,t)]^N. \quad (3)$$

Furthermore, it can be shown (Appendix A) that $[g(z,t) - 1] \propto k_m t/N$ leading to

$$G(z,t) = \lim_{N\to\infty} \exp\{N[g(z,t) - 1]\}. \quad (4)$$

The significance of the above mapping lies in the fact that it exactly maps the original problem [obtaining $G(z,t)$] to a reduced problem [obtaining $g(z,t)$], which is easier to analyze. The simplification provided by this mapping derives from the fact that the number of mRNAs, which is unbounded in the original model, is effectively replaced by a two-state system in the reduced model.

Using Eq. (4), we can readily connect expressions for the mean and Fano factor of the original model

to the corresponding expressions for the reduced model (Appendix A). In particular, we show that the Fano factors for the original and reduced models are identical (in the limit $N \to \infty$). This is a useful result since it is generally easier to obtain the Fano factor for the reduced model.

## IV. EXACT DISTRIBUTIONS FOR THE TWO-STAGE MODEL

We now show how the PPA mapping directly leads to exact results for protein distributions in the two-stage model [Fig. 1(c)]. The two-stage model is the simplest model of stochastic gene expression and has been widely analyzed in both theoretical and experimental studies. While exact results for steady-state distributions have been derived recently [22], the corresponding results for time-dependent distributions have not been obtained so far.

Using the PPA mapping [Fig. 1(c)], we see that the reduced model (obtained by replacing each type of mRNA by a two-state system) for proteins is equivalent to a model for mRNAs with promoter switching. An explicit derivation of the reduced model, starting from the master equation, is provided in Appendix C. The reduced model has been studied in previous work and analytical results for the corresponding mRNA distributions have been obtained [13,14]. Using these results, the generating function for the steady-state distribution of proteins in the reduced model is given by

$$g^*(z) = {}_1F_1\left[\frac{k_m/N}{\mu_p}; \frac{\mu_m}{\mu_p}; \frac{k_p}{\mu_p}(z-1)\right]. \tag{5}$$

Now, using Eq. (4), we obtain that the protein steady-state generating function for the two-stage model is given by

$$G^*(z) = \lim_{N\to\infty} \exp\left(N\left\{{}_1F_1\left[\frac{k_m/N}{\mu_p}; \frac{\mu_m}{\mu_p}; \frac{k_p}{\mu_p}(z-1)\right] - 1\right\}\right). \tag{6}$$

Equation (6), derived directly from known results, is equivalent to the exact result derived recently using a different approach (Appendix C). The concise derivation presented above highlights a general point: the PPA mapping approach leads to protein distributions using results for mRNA distributions for models with promoter-based regulation.

We now apply the PPA mapping to obtain the time-dependent joint distribution of mRNAs and proteins in the original model [with generating function $G(y,z,t)$] using the time-dependent distribution of proteins in the reduced model [with generating function $g(z,t)$]. As noted, the reduced model is equivalent to a model for mRNAs with promoter-based regulation and the corresponding result for the time-dependent generating function of the mRNA distribution has been derived in previous work [15]. Using this result to obtain $g(z,t)$, we derive (Appendix C) that the time-dependent joint generating function of mRNAs and proteins is given by

$$G(y,z,t) = \lim_{N\to\infty} \exp\left\{N\left[g(z,t) + (y-1)\frac{\mu_p}{k_p}\partial_z g(z,t) + \frac{y-1}{k_p(z-1)}\partial_t g(z,t) - 1\right]\right\}. \tag{7}$$

Equation (7) is the most general exact result for the two-stage model of gene expression and all the previously derived results can be obtained from it by taking appropriate limits.

## V. EXACT RESULTS FOR EXTENSIONS OF TWO-STAGE MODEL

### A. Model with multistep mRNA processing

We now show how the partitioning of Poisson processes leads to exact results for some biologically motivated extensions of the two-stage model. Figure 2 presents an extension that allows for an arbitrary number of processing steps for mRNAs. For example, in eukaryotes, these processing steps can represent reactions such as polyadenylation and transport to the cytoplasm, which are required for production of a processed mRNA that is competent for translation. We will call such a processed mRNA a mature mRNA (whereas the unprocessed initial transcript will simply be referred to as a mRNA). Let us now consider the arrival process of a mature mRNA.

The kinetic scheme for the model with $r$ preprocessing steps leading to mature mRNAs is shown in Fig. 2(a). In the following, we invoke the partitioning property of Poisson processes to show that the arrival process of a mature mRNA, in the steady-state limit, is a Poisson process. At any time $t$, we partition the transcribed mRNAs into two types: Type 1 corresponds to a transcribed mRNA that is converted to a mature mRNA by time $t$, and Type 2 includes all the remaining transcribed mRNAs. Let us denote the probability that a transcribed mRNA is classified as Type 1 at time $t$ by $q(t)$. Thus, $q = \lim_{t\to\infty} q(t)$ is the probability that an mRNA transcribed at $t = 0$ is eventually converted into a mature mRNA. Given a mRNA in the $i$th state ($1 \leqslant i \leqslant r-1$), the probability that it is converted into the $(i+1)$th intermediate state without being degraded is $(\frac{k_i}{k_i+\mu_i})$. Thus, in the long-time limit, we have

$$q = \prod_{i=1}^{r}\left(\frac{k_i}{k_i+\mu_i}\right). \tag{8}$$

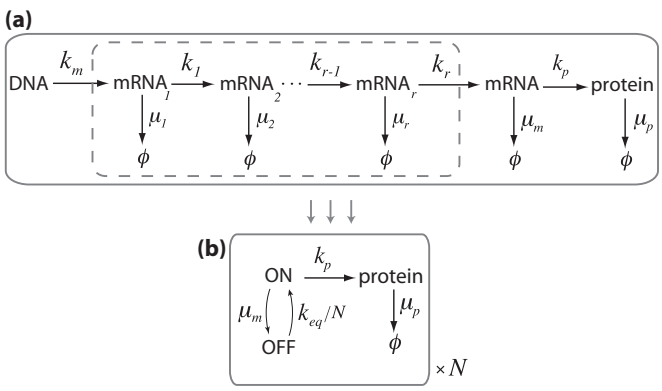


FIG. 2. (a) In this model, mRNAs undergo multistep preprocessing before being competent to produce proteins. Proteins can be created only from the mature mRNA created in the final processing step. (b) Arrival of mature mRNAs is shown to be a Poisson process in steady-state leading to the reduced model shown.

Note that the arrival process of transcribed mRNAs (Type 1 or 2) is a Poisson process with rate $k_m$. In the steady-state limit, the probability that a transcribed mRNA is labeled as Type 1 is $q$. Thus, invoking the partitioning theorem for Poisson processes, we obtain that the arrival process for a Type 1 mRNA (in the steady-state limit) is a Poisson process occurring with rate

$$k_{\text{eq}} = k_m \left(\frac{k_1}{k_1+\mu_1}\right)\ldots\left(\frac{k_r}{k_r+\mu_r}\right). \tag{9}$$

Since a mRNA is classified as Type 1 once it becomes a mature mRNA, it follows that the arrival process of mature mRNAs, in the steady-state limit, is a Poisson process with rate $k_{\text{eq}}$. Some interesting results follow from the preceding observation. First, in the steady-state limit, since mature mRNAs arrive according to a Poisson process, the corresponding reduced model is a two-state model [as in Fig. 1(b)]. Thus, the steady-state distribution of mature mRNAs is a Poisson distribution with mean $k_{\text{eq}}/\mu_m$. Furthermore, the model for proteins is the same as the basic two-stage model [Fig. 1(c)] but with $k_m$ replaced by $k_{\text{eq}}$ [Fig. 2(a)]. Correspondingly, the exact protein steady-state distribution is given by Eq. (6), with the substitution $k_m \to k_{\text{eq}}$. Thus, we obtain that the exact steady-state distribution of proteins for the model in Fig. 2 is given by

$$G(z) = \lim_{N\to\infty} \exp\left(N\left\{{}_1F_1\left[\frac{k_{\text{eq}}/N}{\mu_p};\frac{\mu_m}{\mu_p};\frac{k_p}{\mu_p}(z-1)\right]-1\right\}\right). \tag{10}$$

### B. Model with delayed degradation

The PPA mapping approach can also be applied to models that include non-Markovian processes. An example involving posttranslational regulation leading to a constant delay in the degradation of proteins is illustrated in Fig. 3. The degradation of proteins typically occurs via complex proteolytic pathways involving multiple steps of tagging and binding of auxiliary proteins. A simplified assumption that is commonly used is to replace multistep degradation by a fixed time delay, which motivates the model outlined in Fig. 3. Recent work has analyzed protein steady-state distributions for models with a constant time delay in protein degradation [24–26]. However, the processes of transcription and translation are generally lumped together and it is assumed that proteins are produced in a single step from the DNA in these models. The PPA mapping approach allows us to obtain the exact steady-state protein distributions for a simplified model, which includes both mRNAs and proteins. A detailed derivation (Appendix D) leads to the generating function for arbitrary values of $\tau$. For simplicity, we present here the results in the limit $\tau \ll 1$

$$G^*(z) = \exp\left[\frac{k_m k_p \tau(z-1)}{\mu_m - k_p(z-1)}\right] \lim_{N\to\infty} \times \exp\left(N\left\{{}_1F_1\left[\frac{k_m/N}{\gamma};\frac{\mu_m}{\gamma};\frac{k_p}{\gamma}(z-1)\right]-1\right\}\right). \tag{11}$$

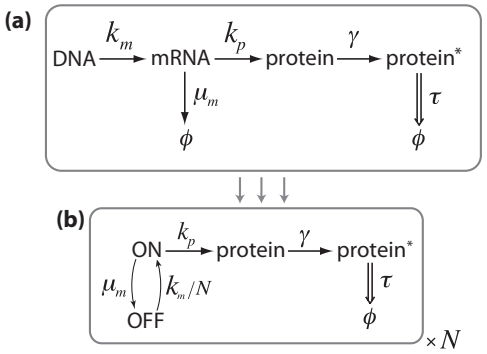


FIG. 3. (a) Kinetic scheme for model with a fixed-time delay in the degradation of proteins. Protein molecules after being tagged (with rate $\gamma$) are degraded after a fixed time delay $\tau$. (b) Mapping of the original model (a) to $N$ independent, identical reduced models ($N \to \infty$).

## VI. DISCUSSION

Several recent experiments have focused on quantifying variations in gene expression and on inference of the underlying mechanisms based on observations of noise [27]. Correspondingly, there is a clear need for theoretical tools to complement such experimental efforts to understand the role of noise in gene expression in diverse cellular processes. The current work addresses this need by developing an analytical framework for obtaining protein distributions for stochastic models of gene expression.

We have shown how the partitioning of Poisson arrival processes can lead to equivalent reduced models that are, in general, simpler to analyze. This mapping can be used to derive exact results for protein distributions using mRNA distributions for models with promoter-based regulation. In recent work, analytical results have been derived for mRNA distributions for a general class of models with promoter-based regulation [16,17]. These results, in combination with the PPA mapping approach developed in this work, can be used to obtain exact protein distributions for a broad class of gene expression models. Furthermore, previous work [28] has shown how a representation using generating functions can be used in developing a variational approach for modeling stochastic cellular processes. Thus, the results obtained in this work, in combination with such variational approaches, can be used to provide quantitative insights into the role of different kinetic schemes in regulating the noise in gene expression.

Noise in gene expression has been shown to play a critical role in diverse cellular processes [1]. It is increasingly becoming clear that quantifying and modeling gene expression variations among single cells in a population can lead to fundamental new insights into old problems. The approach developed in this work can be used to obtain analytical results for multiple extensions of the basic gene expression models. It can be generalized to analyze models including promoter-based regulation, in particular the so-called standard model of gene expression [29]. As more cellular processes are studied using single-cell approaches, the results obtained can guide analysis and interpretation of such experiments. As currently formulated, the approach cannot be used for models with feedback effects (i.e., with rates that depend on protein numbers); however, it is hoped that future work will address

this issue building on current insights. It will also be of interest to extend the PPA mapping approach developed in this work to a broader range of cellular processes for which stochastic effects are critical.

## ACKNOWLEDGMENTS

The authors acknowledge funding support from the NSF through Award No. PHY-0957430. T.P. acknowledges the support of S. Eubank and the NDSSL group at VBI.

## APPENDIX A: CONNECTING ORIGINAL AND REDUCED MODELS

In this section we discuss the relations between the generating functions of the original and reduced models. To begin, we note that the number of mRNAs ($M$) and proteins ($P$) in the original process are, respectively, given by the sum of the number of mRNA ($m$) and protein ($p$) in the $N$ independent and identical reduced processes. We define $\Phi_M(P,t)$ [$\phi_m(p,t)$] as the joint probability to find $M$ ($m$) mRNA and $P$ ($p$) proteins in the original (reduced) process at time $t$. The probability distributions of the original and reduced processes are related via

$$\Phi_M(P,t) = \sum_{m_i,p_i} \delta\left(M - \sum_i m_i\right) \delta\left(P - \sum_i p_i\right) \prod_{i=0}^{N} \phi_{m_i}(p_i,t), \tag{A1}$$

where $\delta(X-Y) = 1$ for $X = Y$ and zero otherwise. It follows that the generating functions, defined by

$$G(y,z,t) = \sum_{M,P} y^M z^P \Phi_M(P,t) \tag{A2}$$

$$g(y,z,t) = \sum_{m,p} y^m z^p \phi_m(p,t) \tag{A3}$$

are related by

$$G(y,z,t) = [g(y,z,t)]^N \tag{A4}$$

as expected for sums of independent and identically distributed random variables. For large $N$ values, successive differentiation shows that the averages in both models are related via

$$\bar{m} = \frac{\bar{M}}{N} \quad \bar{m^2} = \frac{\bar{M^2} - \bar{M}^2}{N}, \tag{A5}$$

$$\bar{p} = \frac{\bar{P}}{N} \quad \bar{p^2} = \frac{\bar{P^2} - \bar{P}^2}{N}. \tag{A6}$$

Correspondingly, the Fano factors for the protein distributions are related by: $F_g = F_G - \bar{P}/N$, so that in the limit $N \to \infty$ $F_g = F_G$, as presented in the main text.

Focussing our attention on the protein distributions, we choose to write $G(z,t) = G(1,z,t)$ and $g(z,t) = g(1,z,t)$. In the following, we consider the limit $N \to \infty$. In this case, up to any time $t$, the production of more than one mRNA by the reduced process is highly unlikely (of second order in $k_m t/N$), as shown in the main text. In the reduced model, one can, therefore, neglect all states with more than one mRNA. Thus, we have

$$g(y,z,t) = g_0(z,t) + y g_1(z,t), \tag{A7}$$

with $g_m(z,t) = \sum_p z^p \phi_m(p,t)$. It follows that

$$g(y,z,t) = g(z,t) + (y-1) g_1(z,t). \tag{A8}$$

In the following, we show that, at the lowest order, the generating function is such that $g(z,t) - 1 \propto k_m t/N$. Let us denote by $\phi_m(p,t|m',p',s)$ the probability distribution at time $t$ with the following condition $\phi_m(p,t=s|m',p',s) = \delta_{m,m'}\delta_{p,p'}$. Since the transition rate from the 0 mRNA state to the 1 mRNA state can be made arbitrarily small ($k_m/N$), we can assume that the system has, at maximum, one transition from the state 0 to 1 (in a given time $t$). Neglecting all events that include more than one transition $0 \to 1$, it follows that $\phi(p,t|0,0,0)$, defined by $\phi_0(p,t|0,0,0) + \phi_1(p,t|0,0,0)$, can be written as

$$\phi(p,t|0,0,0) = \delta(p) e^{-t k_m/N} + \int_0^t ds \frac{k_m}{N} e^{-s k_m/N} \tilde{\phi}(p,t|1,0,s), \tag{A9}$$

where $\exp(-t k_m/N)$ is the probability that we observe no $0 \to 1$ transitions in a time $t$, while $\exp(-s k_m/N) k_m/N ds$ is the probability of a transition between time $s$ and $s + ds$. The distribution $\tilde{\phi}(p,t|1,0,s)$ describes the probability to find $p$ proteins in a process where all transitions $0 \to 1$ are now neglected, and with the condition $m = 1$ and $p = 0$ at time $t = s$. The latter distribution $\tilde{\phi}$, and its generating function $\tilde{g}$, are, therefore, independent of the ratio $k_m/N$. It follows that the generating function $g(z,t)$ [in our case $g(z,t) = g(z,t|0,0,0)$] is

$$g(z,t) = e^{-(k_m/N)t} + \int_0^t ds \frac{k_m}{N} e^{-(k_m/N)s} \tilde{g}(z,t|1,0,s), \tag{A10}$$

which at the first order in $k_m/N$ leads to

$$g(z,t) = 1 + \frac{k_m}{N} \int_0^t ds \, [\tilde{g}(z,t|1,0,s) - 1]. \tag{A11}$$

Using the fact that $\tilde{g}(z,t|1,0,s) = \tilde{g}(z,t-s|1,0,0)$ and defining the dimensionless variable $\alpha = 1 - s/t$ we obtain

$$g(z,t) = 1 + \frac{k_m t}{N} \int_0^1 d\alpha \, [\tilde{g}(z,\alpha t|1,0,0) - 1], \tag{A12}$$

and, thus, $g(z,t) - 1 \propto \frac{k_m t}{N}$ as claimed in the main text.

## APPENDIX B: TWO-STAGE MODEL OF GENE EXPRESSION: MRNA DISTRIBUTION

In this section, we show how the PPA mapping leads to the distribution of mRNA levels for the two-stage model. In Appendix B1, we write down the master equation and define the associated generating function $G(z,t)$. The mapping is then introduced in Appendix B2, by defining the generating function $g(z,t)$ of the reduced model. The time-dependent solution of the reduced process is given in Appendix B3, and, finally, the full generating function $G(z,t)$ is given in Appendix B4.

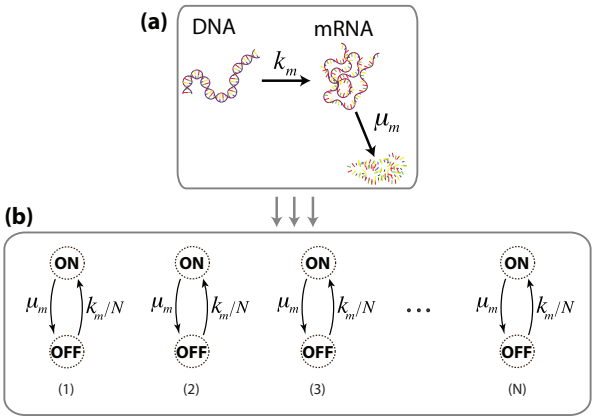


FIG. 4. (Color online) (a) The kinetic scheme for a simple model for mRNA production and decay. (b) Reduced model emerging from the PPA mapping. Probability distribution of number of mRNAs in (a) is identical to the probability distribution of the number of systems in the ON state in (b).

### 1. Master equation and generating function

The master equation for $\Phi_M(t)$, the probability distribution of mRNAs in Fig. 4(a), is given by

$$\partial_t \Phi_M(t) = k_m[\Phi_{M-1}(t) - \Phi_M(t)] + \mu_m[(M+1)\Phi_{M+1}(t) - M\Phi_M(t)]. \quad \text{(B1)}$$

The equation for the generating function $G(z,t) = \sum_M z^M \Phi_M(t)$ is

$$\partial_t G = k_m(z-1)G - \mu_m(z-1)\partial_z G. \quad \text{(B2)}$$

The exact solution can be obtained by directly solving Eq. (B2). However, this problem also provides an ideal example to illustrate the PPA mapping approach.

### 2. Mapping

The PPA mapping connects the original model to $N$ independent, identical reduced models [Fig. 4(b)]. To explicitly derive it from the master equation, let us write the generating function as $G = (g)^N$. Substituting in Eq. (B2), we see that $g$ and $G$ obey the same equation with the rescaling $k_m \to k_m/N$:

$$\partial_t g = \frac{k_m}{N}(z-1)g - \mu_m(z-1)\partial_z g. \quad \text{(B3)}$$

For the reduced model, defining $\phi_m(t)$ as the probability to have $m$ mRNAs at time $t$, we can write the generating function as $g(z,t) = \phi_0(t) + z\phi_1(t) + z^2\phi_2(t)\ldots$. As discussed, for large $N$, it is unlikely to find more than one mRNA in the reduced model. In the stationary state, we have $\phi_0^* \simeq 1 - \mathcal{O}(1/N)$ and $\phi_m^* \simeq \mathcal{O}(1/N^m)$ for $m \geqslant 1$. Keeping the first-order term in $1/N$, the dynamics of the reduced model is effectively described by the kinetic scheme of an ON-OFF model presented in Fig. 4(b).

### 3. The reduced model: Its time-dependent solution

Let us now consider the initial condition $\phi_m(t=0) = \delta_{m,0}$, so that we have $\phi_m(t) \simeq \mathcal{O}(1/N^m)$ for $m \geqslant 1$ and all time $t$. To first order in $1/N$, the generating function of the reduced model is $g(z,t) = \phi_0(t) + z\phi_1(t)$, where $\phi_0(t)$ and $\phi_1(t)$ obey the master equation of the two-state model,

$$\partial_t \phi_0(t) = -\partial_t \phi_1(t) = -\frac{k_m}{N}\phi_0(t) + \mu_m \phi_1(t), \quad \text{(B4)}$$

with solution

$$\phi_1(t) = 1 - \phi_0(t) = (1 - e^{-(\mu_m + k_m/N)t})\phi_1^*, \quad \text{(B5)}$$

where $\phi_1^* = (k_m/N)/(\mu_m + k_m/N)$.

### 4. The full generating function

The full generating function is given by $G = \lim_{N\to\infty}(g)^N = \lim_{N\to\infty} \exp[N(g-1)]$ and leads to

$$G(z,t) = \exp\left[\frac{k_m}{\mu_m}(z-1)(1 - e^{-\mu_m t})\right], \quad \text{(B6)}$$

which corresponds to the well-known Poisson distribution of mRNA, with mean $(k_m/\mu_m)(1 - e^{-\mu_m t})$.

## APPENDIX C: TWO-STAGE MODEL OF GENE EXPRESSION: PROTEIN DISTRIBUTION

In this section we show how the PPA mapping allows us to obtain the protein distribution and the joint mRNA-protein distribution for the two-stage model [Fig. 5(a)]. In Appendix C1, we write down the master equation and define the associated generating function $G(y,z,t)$. Details of the mapping are presented in Appendix C2 by defining the generating function $g(y,z,t)$ of the reduced model. The time-dependent solution of $g(y,z,t)$ is given in Appendix C3, and finally, the full generating function $G(y,z,t)$ is obtained in Appendix C4.

### 1. Master equation and generating function

Let us now consider the full probability distribution of the two-stage model by writing $\Phi_M(P,t)$, the time-dependent probability distribution, with the master equation:

$$\begin{aligned}\partial_t \Phi_M(P,t) = {} & k_m[\Phi_{M-1}(P,t) - \Phi_M(P,t)] \\ & + \mu_m[(M+1)\Phi_{M+1}(P,t) - M\Phi_M(P,t)] \\ & + k_p M[\Phi_M(P-1,t) - \Phi_M(P,t)] \\ & + \mu_p[(P+1)\Phi_M(P+1,t) - P\Phi_M(P,t)].\end{aligned} \quad \text{(C1)}$$

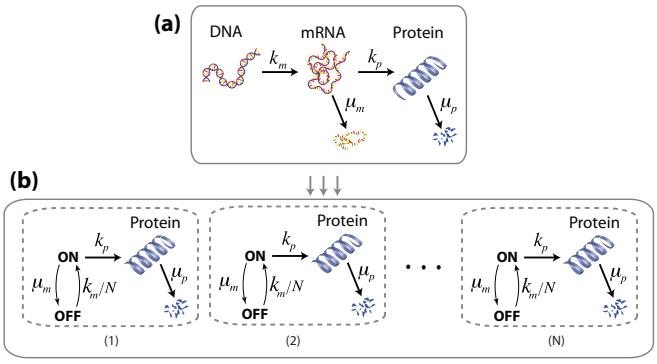


FIG. 5. (Color online) (a) The kinetic scheme for protein production in the two-stage model. (b) Reduced model emerging from the PPA mapping.

The generating function

$$G(y,z,t) = \sum_{M,P} y^M z^P \Phi_M(P,t) \tag{C2}$$

obeys

$$\partial_t G = k_m(y-1)G - \mu_m(y-1)\partial_y G + k_p(z-1)y\partial_y G - \mu_p(z-1)\partial_z G. \tag{C3}$$

### 2. Mapping

Following the steps presented in the previous section, we define $g(y,z,t)$, such that $G = (g)^N$. We see that $g$ is governed by

$$\partial_t g = \frac{k_m}{N}(y-1)g - \mu_m(y-1)\partial_y g + k_p(z-1)y\partial_y g - \mu_p(z-1)\partial_z g. \tag{C4}$$

Again, we see that $g$ corresponds to the generating function of the two-stage model under the rescaling $k_m \to k_m/N$. For large $N$ values, the production of two or more mRNA in the reduced model is unlikely and can be neglected. In the limit $N \to \infty$, the generating function can be written as $g(y,z,t) = \sum_p z^p[\phi_0(p,t) + y\phi_1(p,t)]$. Its dynamics is effectively described by the kinetic scheme presented in Fig. 5(b). Starting with the initial condition $\phi_m(p,t=0) = \delta_{m,0}\delta_{p,0}$, we have $\phi_m(p,t) \simeq 1/N^m$ for $m \geqslant 1$ and $\forall t$.

### 3. The reduced model: Its time-dependent solution

Let us write $g$ in the form $g(y,z,t) = g_0(z,t) + yg_1(z,t)$, where $g_0(z,t)$ and $g_1(z,t)$ are the generating functions defined by $g_m(z,t) = \sum_p z^p\phi_m(p,t)$ $(m = 0,1)$. The latter quantities obey the coupled equations

$$\partial_t g_0 = -\mu_p(z-1)\partial_z g_0 - \frac{k_m}{N}g_0 + \mu_m g_1, \tag{C5}$$

$$\partial_t g_1 = -\mu_p(z-1)\partial_z g_1 + k_p(z-1)g_1 - \mu_m g_1 + \frac{k_m}{N}g_0. \tag{C6}$$

Summing these two equations and writing $g(z,t) = g(1,z,t)$, we get

$$g_1(z,t) = \frac{1}{k_p(z-1)}\partial_t g(z,t) + \frac{\mu_p}{k_p}\partial_z g(z,t), \tag{C7}$$

which allows us to write $g(y,z,t)$ as

$$g(y,z,t) = g(z,t) + (y-1)\frac{\mu_p}{k_p}\partial_z g(z,t) + \frac{(y-1)}{k_p(z-1)}\partial_t g(z,t). \tag{C8}$$

Let us first consider the result for protein distributions in the stationary state. Based on previous work (Refs. [13–15]), we obtain the stationary solution of the reduced model

$$g^*(z,t) = {}_1F_1\left[\frac{k_m/N}{\mu_p}; \frac{\mu_m}{\mu_p}; \frac{k_p}{\mu_p}(z-1)\right], \tag{C9}$$

where ${}_1F_1$ is the confluent hypergeometric function. Furthermore, the time-dependent solution for the protein distribution in the reduced model has been obtained in previous work (Ref. [15]):

$$g(z,t) = F_s(t)\,{}_1F_1\left[\frac{k_m/N}{\mu_p}; \frac{\mu_m}{\mu_p}; \frac{k_p}{\mu_p}(z-1)\right] + F_{ns}(t)\,{}_1F_1\left[1-\frac{\mu_m}{\mu_p}; 2-\frac{\mu_m}{\mu_p}; \frac{k_p}{\mu_p}(z-1)\right], \tag{C10}$$

with

$$F_s(t) = {}_1F_1\left[-\frac{k_m/N}{\mu_p}; 1-\frac{\mu_m}{\mu_p}; -\frac{k_p}{\mu_p}e^{-\mu_m t}(z-1)\right], \tag{C11}$$

$$F_{ns}(t) = \frac{k_m k_p(z-1)}{N\mu_m(\mu_p-\mu_m)}e^{-\mu_m t} \times {}_1F_1\left[\frac{\mu_m}{\mu_p}; 1+\frac{\mu_m}{\mu_p}; -\frac{k_p}{\mu_p}e^{-\mu_m t}(z-1)\right]. \tag{C12}$$

### 4. The full generating function

From $G = (g)^N$, it is readily shown that the original generating function is given by

$$G(y,z,t) = \lim_{N\to\infty} e^{N\mathcal{F}[g(z,t)]}, \tag{C13}$$

with

$$\mathcal{F}[g(z,t)] = g(z,t) + (y-1)\frac{\mu_p}{k_p}\partial_z g(z,t) + \frac{y-1}{k_p(z-1)}\partial_t g(z,t) - 1, \tag{C14}$$

and in the steady-state

$$G^*(y,z) = \lim_{N\to\infty} \exp\left(N\left\{{}_1F_1\left[\frac{k_m/N}{\mu_p}; \frac{\mu_m}{\mu_p}; \frac{k_p}{\mu_p}(z-1)\right] - 1\right\} + (y-1)\frac{k_m}{\mu_m}{}_1F_1\left[1; 1+\frac{\mu_m}{\mu_p}; \frac{k_p}{\mu_p}(z-1)\right]\right). \tag{C15}$$

In the following, we show that the steady-state distribution derived above is equivalent to the exact result derived in recent work (Ref. [22]). By the definition of the hypergeometric functions, we have $\frac{d}{dx}{}_1F_1(\alpha;\beta;\gamma x) = \frac{\alpha}{\beta}\gamma\ {}_1F_1(\alpha+1;\beta+1;\gamma x)$ or ${}_1F_1(\alpha;\beta;\gamma x) = 1 + \frac{\alpha}{\beta}\gamma\int_0^x {}_1F_1(\alpha+1;\beta+1;\gamma s)\,ds$. Using this relation in the preceding equation for $G^*(z)[= G^*(1,z)]$, we obtain

$$G^*(z) = \exp\left\{\frac{k_m k_p}{\mu_m\mu_p}\int_1^z {}_1F_1\left[1;\ 1+\frac{\mu_m}{\mu_p};\ \frac{k_p}{\mu_p}(s-1)\right]ds\right\}, \tag{C16}$$

which is exactly the result derived in previous work (Ref. [22]).

## APPENDIX D: MODEL WITH DELAYED DEGRADATION

We consider an extension of the two-stage model in which the proteins degrade in two steps. First, proteins are tagged

(with rate $\gamma$), and after being tagged they are degraded with a fixed time delay of $\tau$ [Fig. 3(a)]. The corresponding reduced model, obtained using the PPA mapping approach, is shown in Fig. 3(b).

To obtain the exact solution for the steady-state protein distribution, we catergorize the proteins at a given time $t$ (with $t$ large enough such that the system is in steady-state) into two groups: tagged and untagged proteins. Then, at time $t+\tau$, all the tagged proteins will have degraded and the untagged proteins will survive. During the time-interval $\tau$, mRNAs give rise to new proteins that are added to the system. These new proteins will also survive up to time $t+\tau$. Thus, the random variable corresponding to the number of proteins in the system at time $t+\tau$ is the sum of two indepedent random variables: the number of untagged proteins at time $t$ and the number of proteins created in the time interval $[t,t+\tau]$. Let us denote the corresponding generating functions as follows: total proteins $[Q(z)]$, proteins untagged at time $t$ $[U(z)]$, and proteins created in the time interval $[t,t+\tau]$ $[W(z)]$. Since the total number of proteins is the sum of the other two independent random variables, we have $Q(z)=U(z)W(z)$.

The distribution of untagged proteins at time $t$ is the same as the steady-state distribution of proteins in the basic two-stage model (with degradation rate in the basic two-stage model set equal to the tagging rate $\gamma$). The corresponding generating function has been obtained in previous work ([13]) and is given by

$$U(z)=\lim_{N\to\infty}{}_1F_1\left[\frac{k_m}{N\gamma};\frac{\mu_m}{\gamma};\frac{k_p}{\gamma}(z-1)\right]. \tag{D1}$$

Now, we consider the proteins created in the time interval $\tau$. For the reduced model, let $W_0(z)$ and $W_1(z)$ be the generating functions for the protein distribution corresponding to the system being in OFF and ON states, respectively. The following master equations govern the evolution of $W_0(z)$ and $W_1(z)$:

$$\frac{\partial W_0}{\partial t}=-\frac{k_m}{N}W_0+\mu_m W_1 \tag{D2}$$

$$\frac{\partial W_1}{\partial t}=-\frac{k_m}{N}W_0+\mu_m W_1+k_p(z-1)W_1, \tag{D3}$$

therefore,

$$W_1=\frac{1}{k_p(z-1)}\frac{\partial W}{\partial t} \tag{D4}$$

$$W_0=\frac{-1}{k_p(z-1)}\frac{\partial W}{\partial t}+W, \tag{D5}$$

where $W(z)=W_0(z)+W_1(z)$. Correspondingly, we obtain the following equation for $W(z)$:

$$\frac{\partial^2 W}{\partial t^2}+\left[\frac{k_m}{N}+\mu_m-k_p(z-1)\right]\frac{\partial W}{\partial t}-\frac{k_m}{N}k_p(z-1)W=0. \tag{D6}$$

The solution of this ordinary differential equation is given by [13]:

$$W(z,t)=C_1\,e^{[\alpha(z)-\beta(z)]t}+C_2\,e^{[\alpha(z)+\beta(z)]t}, \tag{D7}$$

where $\alpha(z)$ and $\beta(z)$ are

$$2\alpha(z)=k_p(z-1)-\mu_m-\frac{k_m}{N} \tag{D8}$$

$$[2\beta(z)]^2=k_p^2(z-1)^2+2\left(\frac{k_m}{N}-\mu_m\right)k_p(z-1)+\left(\mu_m+\frac{k_m}{N}\right)^2. \tag{D9}$$

To obtain $C_1$ and $C_2$ we use the initial conditions. Since we are in the steady-state limit, the initial conditions are

$$W_0(z,0)=\frac{\mu_m}{\frac{k_m}{N}+\mu_m}=1-\frac{k_m}{N\mu_m},\quad W_1(z,0)=\frac{k_m}{N\mu_m}. \tag{D10}$$

Using the above, we get

$$C_1=\frac{[\beta(z)+\alpha(z)]-k_p(z-1)W_1(0)}{2\beta(z)} \tag{D11}$$

$$C_2=\frac{[\beta(z)-\alpha(z)]+k_p(z-1)W_1(0)}{2\beta(z)}. \tag{D12}$$

For $N\to\infty$ and $t=\tau$,

$$W(z,\tau)=1+\frac{1}{N}\frac{k_m k_p}{\mu_m^2}\frac{(z-1)}{1-\frac{k_p}{\mu_m}(z-1)}\left(\mu_m\tau-\frac{k_p}{\mu_m}\times\frac{(z-1)}{1-\frac{k_p}{\mu_m}(z-1)}\left\{1-e^{-\mu_m[1-\frac{k_p}{\mu_m}(z-1)]\tau}\right\}\right). \tag{D13}$$

The generating function of the original model is $G(z)=\lim_{N\to\infty}Q^N$:

$$G(z)=\exp\left(\frac{k_m}{\mu_m}\frac{k_p(z-1)}{s(z)}\left\{\mu_m\tau-\frac{k_p(z-1)}{s(z)}[1-e^{-s(z)\tau}]\right\}\right)\times\lim_{N\to\infty}\exp\left(N\left\{{}_1F_1\left[\frac{k_m/N}{\gamma};\frac{\mu_m}{\gamma};\frac{k_p}{\gamma}(z-1)\right]-1\right\}\right), \tag{D14}$$

where $s(z)=\mu_m-k_p(z-1)$.


[1] G. Balázsi, A. van Oudenaarden, and J. Collins, Cell **144**, 910 (2011).

[2] A. Raj and A. van Oudenaarden, Cell **135**, 216 (2008).

[3] O. Gefen, C. Gabay, M. Mumcuoglu, G. Engel, and N. Balaban, Proc. Natl. Acad. Sci. USA **105**, 6145 (2008).

[4] L. Weinberger, J. Burnett, J. Toettcher, A. Arkin, and D. Schaffer, Cell **122**, 169 (2005).

[5] J. Raser and E. O'Shea, Science **304**, 1811 (2004).

[6] J. Paulsson, Phys. Life Rev. **2**, 157 (2005).

[7] Á. Sánchez and J. Kondev, Proc. Natl. Acad. Sci. USA **105**, 5081 (2008).

[8] A. Coulon, O. Gandrillon, and G. Beslon, BMC Syst. Biol. **4**, 2 (2010).

[9] A. Singh, B. Razooky, R. Dar, and L. Weinberger, Mol. Syst. Biol. **8**, 607 (2012).
[10] T. To and N. Maheshri, Science **327**, 1142 (2010).
[11] Y. Taniguchi, P. Choi, G. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. Xie, Science **329**, 533 (2010).
[12] M. Ferguson, D. Le Coq, M. Jules, S. Aymerich, O. Radulescu, N. Declerck, and C. Royer, Proc. Natl. Acad. Sci. USA **109**, 155 (2012).
[13] J. Peccoud and B. Ycart, Theoret. Pop. Biol. **48**, 222 (1995).
[14] A. Raj, C. Peskin, D. Tranchina, D. Vargas, and S. Tyagi, PLoS Biol. **4**, e309 (2006).
[15] S. Iyer-Biswas, F. Hayot, and C. Jayaprakash, Phys. Rev. E **79**, 031911 (2009).
[16] J. Zhang, L. Chen, and T. Zhou, Biophys. J. **102**, 1247 (2012).
[17] A. Stinchcombe, C. Peskin, and D. Tranchina, Phys. Rev. E **85**, 061919 (2012).
[18] J. E. M. Hornos, D. Schultz, G. C. P. Innocentini, J. Wang, A. M. Walczak, J. N. Onuchic, and P. G. Wolynes, Phys. Rev. E **72**, 051907 (2005).
[19] A. F. Ramos, G. C. P. Innocentini, and J. E. M. Hornos, Phys. Rev. E **83**, 062902 (2011).
[20] N. Friedman, L. Cai, and X. Xie, Phys. Rev. Lett. **97**, 168302 (2006).
[21] V. Shahrezaei and P. Swain, Proc. Natl. Acad. Sci. USA **105**, 17256 (2008).
[22] P. Bokes, J. King, A. Wood, and M. Loose, J. Math. Biol. **64**, 829 (2012).
[23] S. Ross, *Introduction to Probability Models* (Academic Press, New York, 2006).
[24] D. Bratsun, D. Volfson, L. Tsimring, and J. Hasty, Proc. Natl. Acad. Sci. USA **102**, 14593 (2005).
[25] L. Lafuerza and R. Toral, Phys. Rev. E **84**, 051121 (2011).
[26] J. Miękisz, J. Poleszczuk, M. Bodnar, and U. Foryś, Bull. Math. Biol. **73**, 2231 (2011).
[27] B. Munsky, G. Neuert, and A. van Oudenaarden, Science **336**, 183 (2012).
[28] Y. Lan, P. G. Wolynes, and G. A. Papoian, J. Chem. Phys. **125**, 124106 (2006).
[29] D. Larson, Curr. Opin. Genet. Dev. **21**, 591 (2011).